Behaviour analysis of Users from Database Request Logs

Interim Report

Gan Wen Xuan Damien
Lewis Poh Yu Gui
Louis Tan Kai Huang

Contents

1.0 Overview

SMU Libraries provide academic materials to SMU community through various means and one particular approach is through the provision of materials via databases. For certain databases, SMU Libraries implemented the Patron Driven Acquisition (PDA) scheme whereby e-books are automatically purchased after meeting certain business rules in placed which aims to quantify the demands of user. The aim of PDA scheme is to ensure that materials purchased are widely utilized across SMU community and improve upon traditional methods of purchase through librarians.

1.1 Business Problem

Currently, funds allocated to PDA scheme are being utilized quickly and a preliminary investigation was conducted which shows that many of such purchases were triggered by a single user via multiple access or download. This defeats the purpose for which PDA scheme is implemented for and has resulted in wastage of funds. As a result, the PDA scheme of some databases are being halt at the moment.

1.2 Objectives

Our team aims to analyse these databases and verify this phenomenon. Subsequently, we will delve deeper into 2 non-PDA databases to obtain more insights into user behaviour patterns to determine the viability of the PDA scheme and propose more efficient business rules to better quantify the demands of users.

2.0 Datasets

Current datasets given are the proxy log data, user master list and Jstor's PDA book usage report.

Proxy Log Data

Proxy log data records each user access to various academic materials provided by SMU Libraries via the libraries' proxy. Every click or download via SMU Libraries' portal will be recorded and identified by the 'Session ID' and 'User ID'.

User Master List

User master list shows the complete list of users in SMU community and is identified by their unique email. More detailed information such as faculty and admission year are also provided where relevant.

Jstor's PDA book usage report

Jstor's PDA book usage report shows the complete list of e-books under PDA scheme. The table also contains information on purchased date and prices for the corresponding books if applicable.

2.1 Issues with Datasets

Complexity of logs files

The URL column of proxy log data holds the most amount of meaningful information such as 'Doc ID' and 'Chapter' that is required for our analysis. However, URL is a long string from which we had to manually identify the position whereby the information is at amongst the noise as seen below.

http://www.jstor.org:80/stable/10.3138/j.ctt5hjwqr.110?Search=yes&resultItemClick=true&searchText=conscription&searchUri=%2Faction%2FdoBasicSearch%3Ffacet_chapter%3DY2hhcHRlcg%253D%253D%26amp%3Bgroup%3Dnone%26amp%3Bwc%3Don%26amp%3Bpage%3D1%26amp%3Bfc%3Doff%26amp%3BQuery%3Dconscription%26amp%3Bacc%3Don%26amp%3Bed%3D2017%26amp%3BsearchType%3DfacetSearch%26amp%3Bsd%3D2010%26amp%3BresultsServiceName%3DdoBackToBasicResults

Multiple Variation of URLs in URL columns

A particular e-book can be access by user through various means such as manual search or browsing, this promotes some variability within the URL recorded in the proxy log data. Moreover, users are given the options between online view browsing and pdf download which further complicates the URL. An example of an e-book having 2 different URLs is show below,

http://www.jstor.org:80/tc/accept?origin=/stable/pdf/j.ctt1bkm5nd.109.pdf?refreqid=search%3A5998b4652239054161320 78f1aed83e5&_=1505450011690

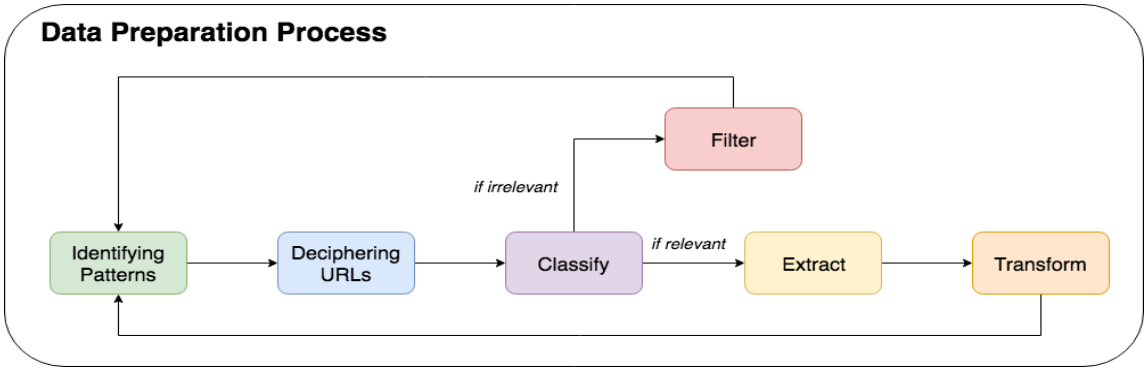
http://www.jstor.org:80/stable/j.ctt1bkm5nd.109?Search=yes&resultItemClick=true&searchText=propaganda&searchText=model&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dpropaganda%2Bmodel&refreqid=search%3A5998b465223905416132078f1aed83e5

Figuring out how vendors assess trigger points attributes

Business rules that trigger the automatic purchase under PDA scheme is either 6 chapter views (online) or 4 chapters download for Jstor database. In Jstor's PDA book usage report, there are attributes corresponding these business rules which are recorded by the vendor. However, how the corresponding chapter views and downloads captured by the vendor is reflected in the proxy log data given by SMU Libraries is not known. As such, more investigation will have to be conducted before further analysis can continue.

3.0 Data Preparation

SMU Libraries provided us data logs with metadata, but minimal information was given regarding the information available within the URL column, which consists of many information needed for our analysis. Our observations found out that the URL column consists of information about database, type of document, document ID and chapter of e-book. When accessing each database, the URL generated for each page differs. As each database URL generated is different in structure and some information may be available for a database but lacking in another, resulting in a need for different data retrieval method.



In preparing the data, a 5-step systematic approach was formulated which consists of identifying URL patterns, deciphering the URL, classifying it based on relevancy, extracting data or filtering it out and lastly, transforming data to suit our analytical needs.

3.1 Identifying Patterns

With the use of software, Charles Proxy, we attempt to replicate how the proxy log data is generated by browsing the database with the 'Doc ID' taken from the dataset and clicking on the various links available on the e-book. In doing so, we form expectations on the various interactions and map to the type of URL generated.

When identifying common patterns in URL, the commonly used directory will be identified.

Non-Material Rows

http://www.jstor.org:80/assets/article-view_20170628T1402/build/article-view/css/article-view.css
http://www.jstor.org:80/assets/article-view_20170628T1402/build/images/favicon.ico
http://www.jstor.org:80/assets/global_20170628T1500/build/global/js/global.min.js
http://www.jstor.org:80/assets/global_20170628T1500/build/images/MagnifyingGlass_22px.svg?1498676380

Amongst the non-material rows identified, for the example above, '/assets/' has been identified to be a common directory used.

Material Rows

http://www.jstor.org:80/stable/10.3138/j.ctt5hjwqr.110?Search=yes&resultItemClick=true&searchText=conscription&searchUri=%2Faction%2FdoBasicSearch%3Ffacet_chapter%3DY2hhcHRlcg%253D%253D%26amp%3Bgroup%3Dnone%26amp%3Bwc%3Don%26amp%3Bpage%3D1%26amp%3Bfc%3Doff%26amp%3BQuery%3Dconscription%26amp%3Bacc%3Don%26amp%3Bed%3D2017%26amp%3BsearchType%3DfacetSearch%26amp%3Bsd%3D2010%26amp%3BresultsServiceName%3DdoBackToBasicResults
http://www.jstor.org:80/stable/pdf/10.3366/j.ctt1g0b6rb.30.pdf?refreqid=search%3A0115189cc95abd84f629fb08ceeb7e5b
http://www.jstor.org:80/stable/pdf/10.2307/j.ctt1gxpcnv.27.pdf
http://www.jstor.org:80/stable/10.2307/j.ctt5hk0m2.20?Search=yes&resultItemClick=true&searchText=alcohol&searchText=and&searchText=reaction&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dalcohol%2Band%2Breaction

For the material rows identified above, '/stable/' is a common directory identified. But it differs because users are given the option to download the document or viewing the material online. For online view, the identifier would be the directory after '10.XXXX/' whereas for downloads, the identifier would be after '/pdf/10.XXXX/'.

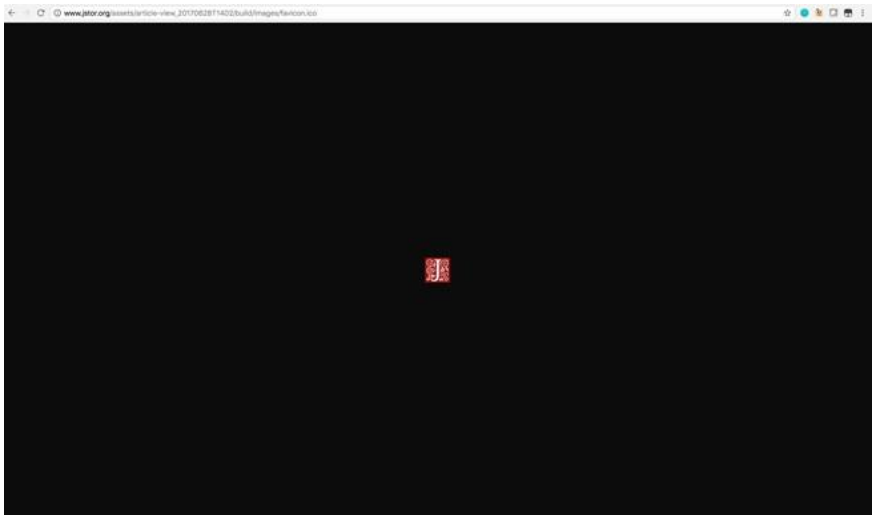
3.2 Deciphering URLs

For each of the sample URL extracted, it will be checked on the browser to see what the page loads.

Non-material Row

```
http://www.jstor.org:80/assets/article-
view_20170628T1402/build/images/favicon.ico
```

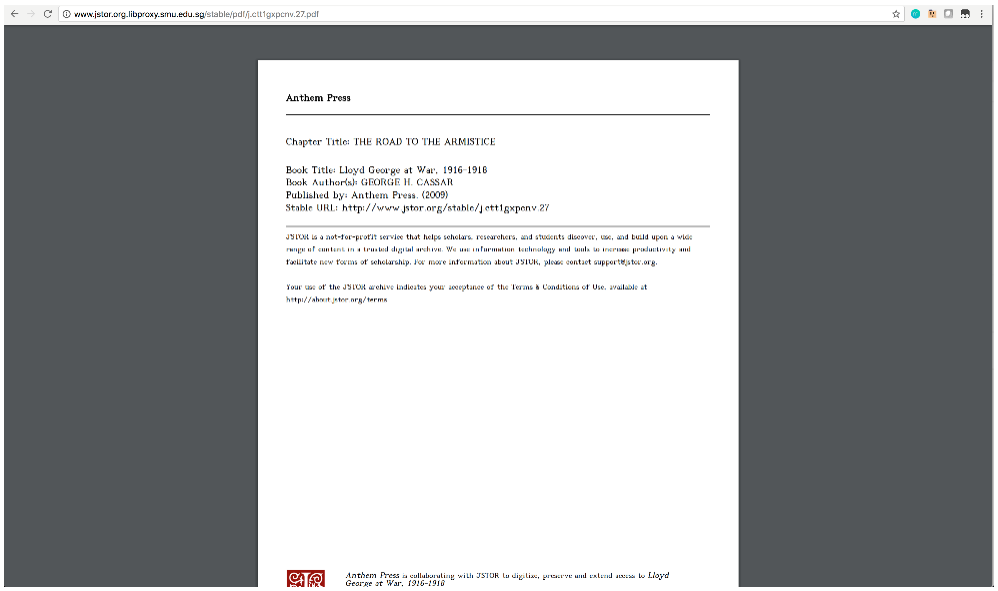
Non-material Row Result



Material Row

```
http://www.jstor.org:80/stable/pdf/10.2307/j.ctt1gxpcnv.27.pd
f
```

Material Row Result



3.3 Classifying

Judging on the content displayed, the URL accessed in the log will be deemed relevant for our analysis or not.

For the non-material row result above, the logo of Jstor is displayed, suggesting that the '/assets/' directory holds the images and styling files (CSS, JavaScript), which we would not need and therefore be classified as irrelevant.

As for the material row result above, the content shown is the actual pdf file of the e-book which is useful for our analysis.

3.4 Extract or Filter

Based on the data relevancy, we will choose to extract the information to be a specific column or filter the entire record out of our analysis.

In the instance of non-material rows, since all the URL with '/assets/' directory are not needed in our analysis, it will be programmatically filtered out. While the material rows will be extracted for further transformation and analysis.

3.5 Transformation

Data like timestamp will need to be transformed to suit the SQL's standard of storing datatype datetime so that we can use functions for the datetime stored. Once data transformation is complete, sample checks are conducted on the cleaned data to ensure that the data is coherent before moving on to undergo exploratory data analysis (EDA).

3.6 Cleaned Data

After performing the data cleaning process, the metadata of extracted data is given below,

Proxy Logs Data

Name	Data type	Length	Description
IP Address	VARCHAR	15	Internet Protocol address of users
Session ID	CHAR	15	Unique Session identifier of users
User ID	CHAR	64	Email of users which is hashed for privacy reasons
URL	VARCHAR	250	URL access by users for the log
Doc ID	VARCHAR	20	Unique Identifier of file accessed by users
DateTime	DATETIME	17	Date and Time at point of access. Format is DD-MMM-YYYY HH:MM:SS e.g '05Dec2017 10:28:24'
Downloads	BOOLEAN	1	Binary attribute whereby 0 is online view and 1 is pdf download
Chapter	NUMERIC	2	Specific chapter access within the e-book

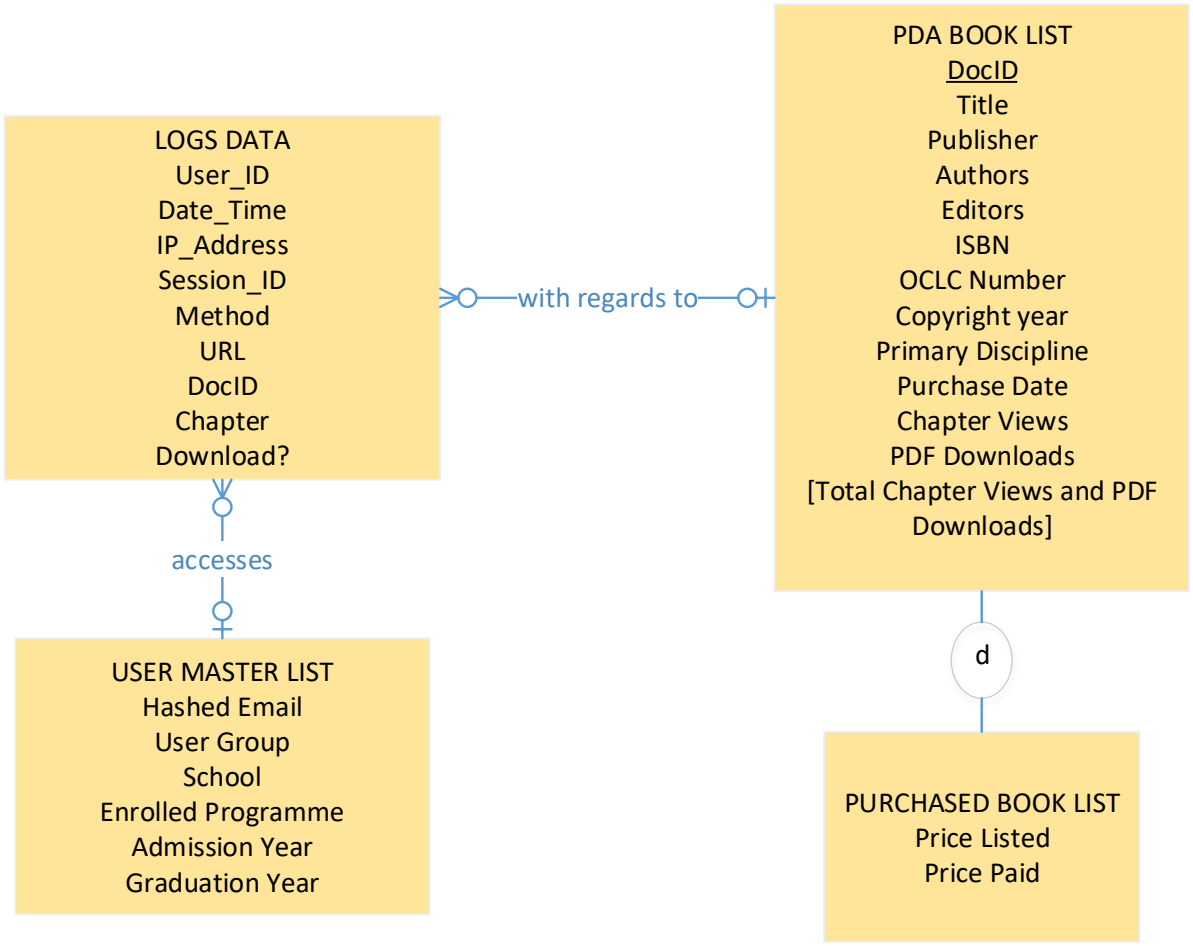
PDA Book List

Name	Data type	Length	Description
Title	VARCHAR	250	Title of book
Doc ID	VARCHAR	20	Unique Identifier of PDA e-book
Publisher	VARCHAR	50	Publisher of e-book
Author	VARCHAR	50	Author of e-book
Editor	VARCHAR	50	Co-Editor of e-book (Can be null)
ISBN	NUMERIC	13	International standard book number
Primary Discipline	VARCHAR	20	Genre of e-book
Purchase Date	DATETIME	17	Purchase date of e-book (can be null) Format is DD/MM/YY HH:MM e.g '12/5/17 00:00'
Chapter View	NUMERIC	2	Number of chapters viewed
PDF Download	NUMERIC	2	Number of chapters downloaded
Listed Price	NUMERIC	4	Listed price of e-book
Price Paid	NUMERIC	4	Price paid for e-book (usually at a discount)

User Master List

Name	Data Type	Length	Description
User ID	CHAR	64	Contains the email of the user
User Group	VARCHAR	50	Contain the category that user belongs
School	VARCHAR	50	Contain school which user belongs
Area of Study	VARCHAR	100	Programme that user is enrolled under
Admission Year	CHAR	7	Admission year of user
Graduation year	CHAR	7	Graduation or expected graduation year

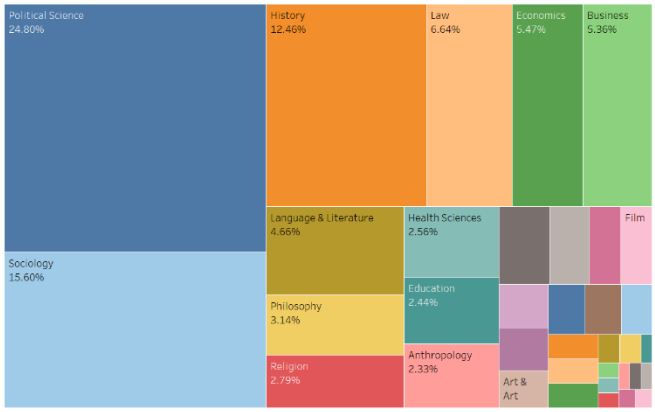
To see how the datasets interact with each other, we map out a Entity-Relationship Diagram for easy visualization. The Entity-Relationship Diagram is as follows,

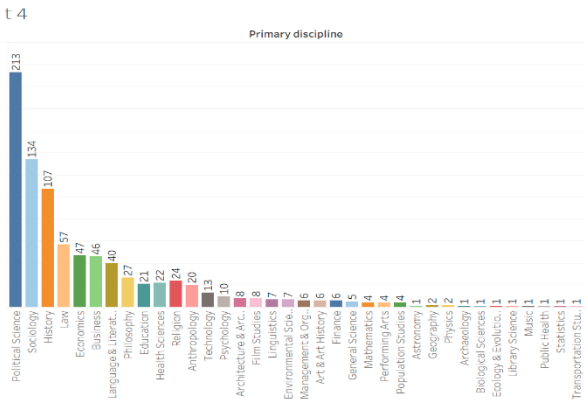


4.0 Exploratory Data Analysis

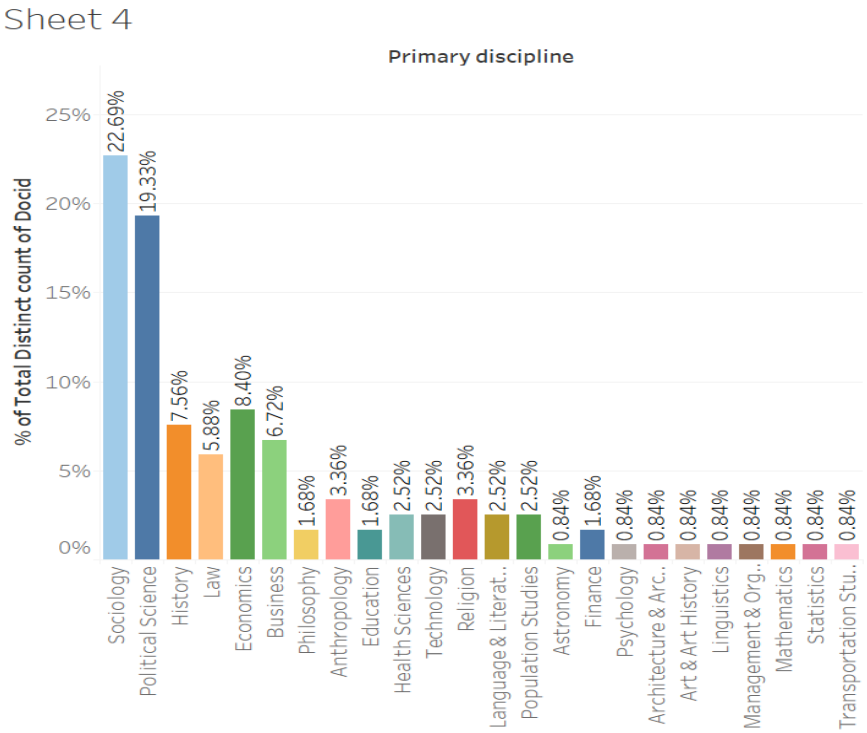
4.1 Profiling of PDA e-books in JSTOR

<PDA Books by Genre>

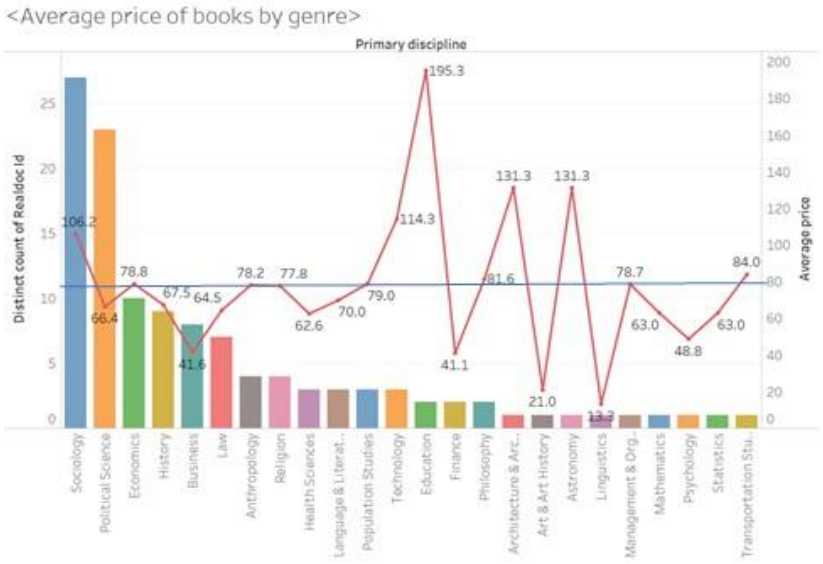




By visualizing the breakdown of books under PDA scheme, we noted that that most books listed under PDA corresponds to liberal acts subjects such as 'Political Science', 'History' and 'Sociology', constituting more than 50% of the total books.

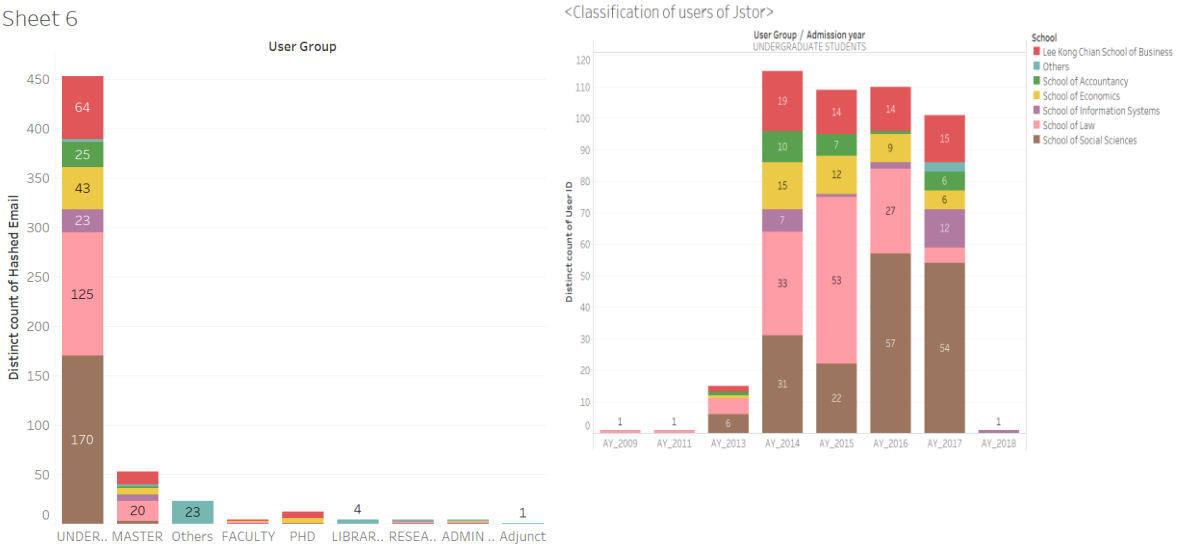


The genre proportions of e-books under PDA subsequently affects the proportion of genre of e-books being purchase. However, while 'Political Science' books held the greatest volume under total books listed, 'Sociology' books are the most purchased.



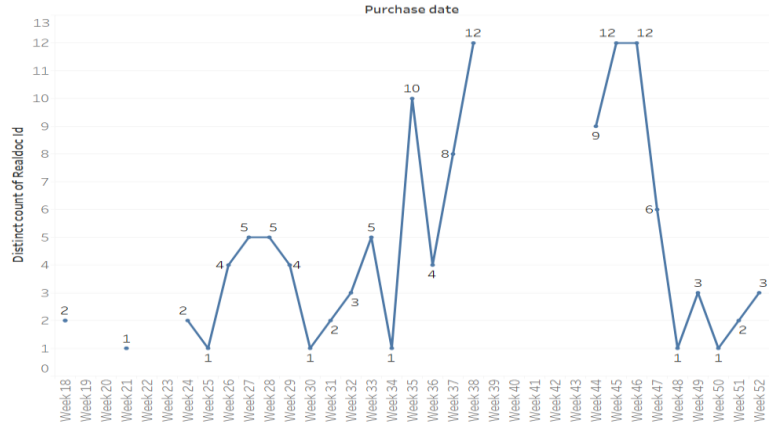
Recognizing that both volume and prices has an impact on how quickly PDA funds are being utilized, we plot a combination chart and note that the total average prices (blue line) is at $77.8. From here we noted that 'Sociology' books are the consuming most of the PDA funds since its volume and average price is well above average.

4.2 Profiling Users of JSTOR



A breakdown of users of JSTOR shows that most users are 'UNDERGRADUATE STUDENTS' and most belonging to either Law or Social Science faculty. This is reasonable because most of the books listed in JSTOR would correspond to the interest of such group of users.

<Time of purchases>

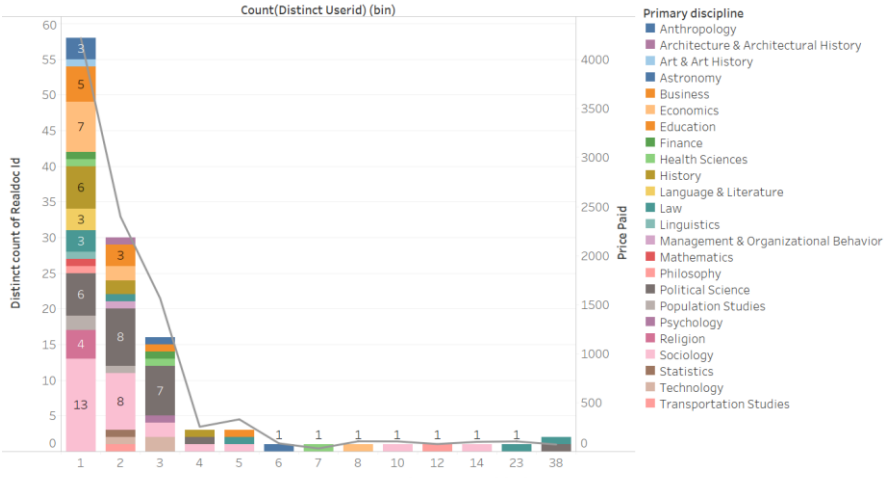


The trend of distinct count of Realdoc Id for Purchase date Week. The marks are labeled by distinct count of Realdoc Id. The data is filtered on Count(Distinct Userid) (bin) and Purchases. The Count(Distinct Userid) (bin) filter keeps 13 of 13 members. The Purchases filter keeps 1.

Looking at the timing of purchases in a time-series diagram, we see that most of the purchases are done from week 35 to week 47 which corresponds to the Semester 1 of regular school term. More specifically, most of the books are purchase during week 37, 45 and 46 which are week 5,12 and 13 of school term. Hence it is apparent that these books are mostly used for research purposes for project basis.

4.3 Book Utilization Rate

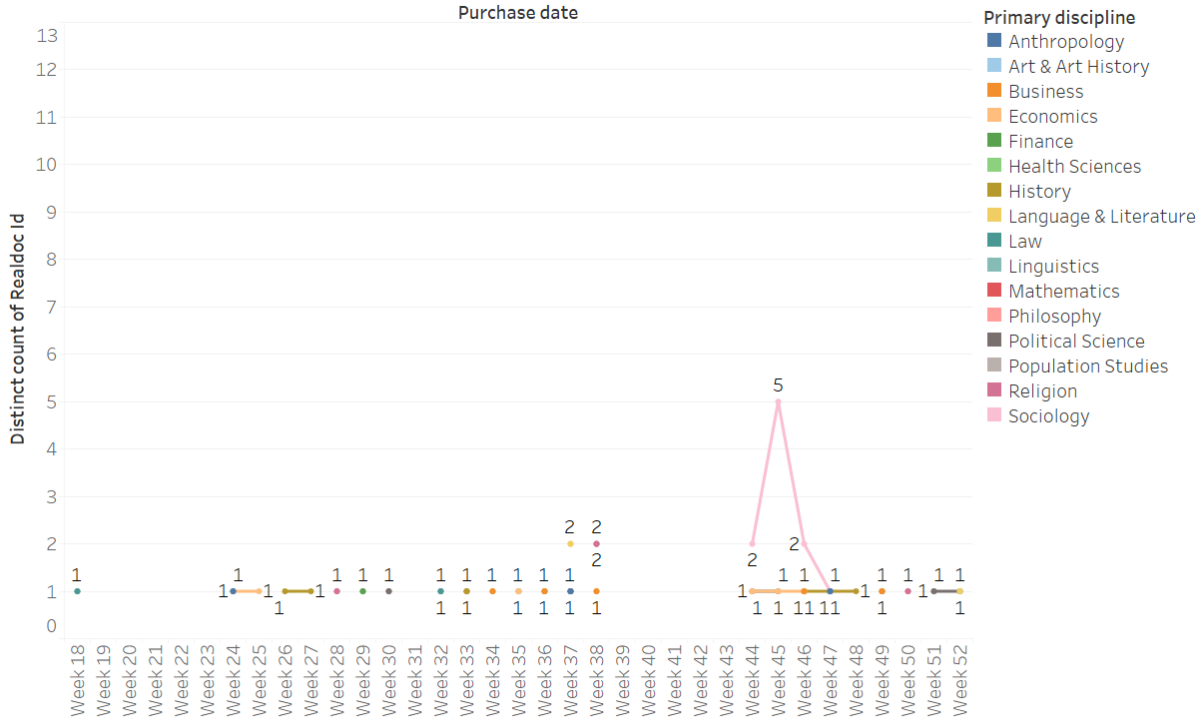
<Book Utilization Breakdown>



The trends of distinct count of Realdoc Id and sum of Price Paid for Count(Distinct Userid) (bin). For pane Distinct count of Realdoc Id: Color shows details about Primary discipline. The marks are labeled by distinct count of Realdoc Id. The data is filtered on Purchases, which keeps 1.

Breaking down into the book utilization rate, we confirm that most of the book purchased are indeed only being utilized by 1 particular user. Only a small number of books purchased via PDA scheme has been widely use. This confirms the preliminary investigation that PDA scheme is not effective in determining assessing the needs of SMU community.

<Time of purchases of e-books used by 1 user only>



The trend of distinct count of Realdoc Id for Purchase date Week. Color shows details about Primary discipline. The marks are labeled by distinct count of Realdoc Id. The data is filtered on Count(Distinct Userid) (bin) and Purchases. The Count(Distinct Userid) (bin) filter keeps 1. The Purchases filter keeps 1.

When we look at time series chart for books used only by 1 user, we also see that most of it is being purchased at Week 45 (Week 12 of term). Hence this confirms that most users could be downloading/viewing books for trivial purposes or unable to search effectively for their academical needs.

5.0 Moving Forward

After confirming our initial expectations and obtain a broad understanding data based on users and genre of e-books, we will begin our analysis user behaviour patterns based on non-PDA databases such as ACLS and EBSCO. Our analysis will cover topics such as chapter views that users tend to download while browsing through academic materials.

From there, we seek to propose new trigger points to which we will attempt to test these proposed business rules from the current JSTOR dataset and record the effects.